

Mixed Cumulative Distribution Networks

Ricardo Silva

ricardo@stats.ucl.ac.uk

Department of Statistical Science University College London

Charles Blundell

c.blundell@gatsby.ucl.ac.uk

Gatsby Computational Neuroscience Unit
University College London

Yee Whye Teh

ywteh@gatsby.ucl.ac.uk

Gatsby Computational Neuroscience Unit
University College London

September 1, 2010

Abstract

Directed acyclic graphs (DAGs) are a popular framework to express multivariate probability distributions. Acyclic directed mixed graphs (ADMGs) are generalizations of DAGs that can succinctly capture much richer sets of conditional independencies, and are especially useful in modeling the effects of latent variables implicitly. Unfortunately there are currently no good parameterizations of general ADMGs. In this paper, we apply recent work on cumulative distribution networks and copulas to propose one general construction for ADMG models. We consider a simple parameter estimation approach, and report some encouraging experimental results.

1 Contribution

Graphical models provide a powerful framework for encoding independence constraints in a multivariate distribution [17, 14]. Two of the most common families, the directed acyclic graph (DAG) and the undirected network, have complementary properties. For instance, DAGs are non-monotonic independence models, in the sense that conditioning on extra variables can also destroy independencies (sometimes known as the “explaining away” phenomenon [17]). Undirected networks allow for flexible “symmetric” parameterizations that do not require a particular ordering of the variables.

More recently, alternative graphical models that allow for both directed and symmetric relationships have been introduced. The *acyclic directed mixed graph* (ADMG) has both directed and bi-directed edges and it is the result of *marginalizing* a DAG: Figure 1 provides an example. [21, 19] show that DAGs are not closed under marginalization, but ADMGs are. Reading off independence constraints from a ADMG can be done with a procedure essentially identical to d-separation [17, 21].

Theoretical properties and practical applications of ADMGs are further discussed in detail by e.g. [2, 25, 5, 29, 18, 24, 10]. One can also have latent variable ADMG models, where bi-directed edges represent a subset of latent variables that have been marginalized. In sparse models, using bi-directed edges in ADMGs frees us from having to specify exactly which latent variables exist and how they might be connected. In the context of Bayesian inference, Markov chain Monte Carlo in ADMGs might have much better mixing properties compared to models where all latent variables are explicitly included [24].

However, it is hard in general to parameterize a likelihood function that obeys the independence constraints encoded in an ADMG. Gaussian likelihood functions and their variations (e.g., mixture models and probit models) have been the only families exploited in most of the literature [21, 24]. The contribution of this paper is to provide a flexible construction procedure to design probability mass functions and density functions that are Markov with respect to

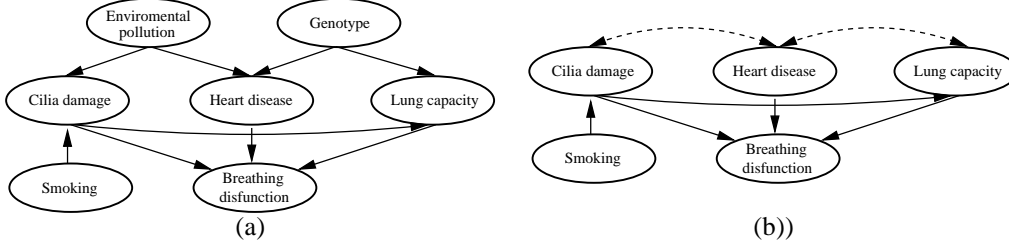


Figure 1: (a) A DAG representing dependencies over a set of variables (adapted from [25], page 137) in a medical domain. (b) The ADMG representing conditional independencies corresponding to (a), but only among the remaining vertices: pollution and genotype factors were marginalized. In general, bi-directed edges emerge from unspecified variables that have been marginalized but still have an effect on the remaining variables. The ADMG is acyclic in the sense that there are no cycles composed of directed edges only. In general, a DAG cannot represent the remaining set of independence constraints after some variables in another DAG have been marginalized.

an arbitrary ADMG. This is done by exploiting recent work on *cumulative distribution networks* [9] and *copulas* [16, 13]. We also provide a straightforward approach to learning in our ADMGs inspired by the parameter estimation approaches in the copula literature. We review mixed graphs and cumulative distribution networks in Section 2. The full formalism is given in detail in Section 3. An instantiation of the framework based on copulas and a parameter estimation procedure is described in Section 4. Experiments are described in Section 5, and we conclude with Section 6.

2 Mixed Graphs and Cumulative Distribution Networks

In this section, we provide a summary of the relevant properties of mixed graph models and cumulative distribution networks, and the relationship between formalisms.

A *bi-directed* graph is a special case of a ADMG without directed edges. The absence of an edge (X_i, X_j) implies that X_i and X_j are *marginally independent*. Hence, bi-directed models are *models of marginal independence* [5]. Just like in a DAG, conditioning on a vertex that is the endpoint of two arrowheads will make some variables dependent. For instance, for a bi-directed graph $X_1 \leftrightarrow X_2 \leftrightarrow X_3$, we have that $X_1 \perp\!\!\!\perp X_3$ but $X_1 \not\perp\!\!\!\perp X_3 | X_2$. See [4, 5] for a full discussion.

Current parameterizations of bi-directed graphs suffer from a number of practical difficulties. For example, consider binary bi-directed graphs, where a complete parameterization was introduced by Drton and Richardson [5]. Let \mathcal{G} be a bi-directed graph with vertex set X_V . Let $q_A \equiv P(X_A = 0)$, for any vertex set X_A contained in X_V . The joint probability $P(X_A = 0, X_{V \setminus A} = 1)$ is given by

$$P(X_A = 0, X_{V \setminus A} = 1) = \sum_{B: A \subseteq B} (-1)^{|B \setminus A|} q_B \quad (1)$$

The set $\{q_S : S \subset S\}$ is known as the Möbius parameterization of $P(X_V)$, since relationship (1) is an instance of the Möbius inversion operation [14]. The marginal independence of the bi-directed graph implies $P(X_A = 0, X_B = 0) = P(X_A = 0)P(X_B = 0)$ if no element in X_A is adjacent to any element in X_B in \mathcal{G} . Therefore, the set of independent parameters in this parameterization is given by $\{q_A\}$, for all X_A that forms a connected set in \mathcal{G} . This parameterization is complete, in the sense that *any* binary model that is Markov with respect to \mathcal{G} can be represented by the set $\{q_A\}$. However, this comes at a price: in general, the number of connected sets can grow exponentially in $|X_V|$ even for a sparse, tree-structured, graph. Moreover, the set $\{q_A\}$ is not *variation independent* [14]: the parameter space is defined by exponentially many constraints. In contrast, different conditional probability tables in a given Bayesian network can be parameterized independently [14, 17].

Cumulative distribution networks (CDNs), introduced by Huang and Frey [9] as a convenient family of cumulative distribution functions (CDFs), provide a alternative construction of bi-directed models by indirectly introducing additional constraints to reduce the total number of parameters. Let X_V be a set of random variables, and let \mathcal{G} be a

bi-directed graph¹ with \mathcal{C} being a set of cliques in \mathcal{G} . The CDF over X_V is given by

$$P(X_V \leq x_V) \equiv F(x_V) = \prod_{S \in \mathcal{C}} F_S(x_S) \quad (2)$$

where each F_S is a parametrized CDF over X_S . A sufficient condition for (2) to define a valid CDF is that each F_S is itself a CDF. CDNs satisfy the conditional independence constraints of bi-directed graphs [9]. For example, consider $X_1 \leftrightarrow X_2 \leftrightarrow X_3$, with cliques $X_{S_1} = \{X_1, X_2\}$ and $X_{S_2} = \{X_2, X_3\}$. The marginal CDF of X_1 and X_3 is $P(X_1 \leq x_1, X_3 \leq x_3) = P(X_1 \leq x_1, X_2 \leq \infty, X_3 \leq x_3) = F_1(x_1, \infty)F_2(\infty, x_3)$. Since this factorizes, it follows that X_1 and X_3 are marginally independent.

The relationship between the complete parameterization of Drton and Richardson and the CDN parameterization can be exemplified in the discrete case. Let each X_i take values in $\{0, 1, 2, \dots\}$. Recall that the relationship between a CDF and a probability mass function is given by the following inclusion-exclusion formula [12]:

$$P(x_1, \dots, x_d) = \sum_{z_1=0}^1 \sum_{z_2=0}^1 \dots \sum_{z_d=0}^1 (-1)^{z_1+z_2+\dots+z_d} F(x_1-z_1, x_2-z_2, \dots, x_d-z_d), \quad (3)$$

for $d = |X_V|$. In the binary case, since $q_A = P(X_A = 0) = P(X_A \leq 0, X_{V \setminus A} \leq 1) = F(x_A = 0, x_{V \setminus A} = 1)$, one can check that (3) and (1) are the same expression. The difference between the CDN parameterization [9] and the complete parameterization [5] is that, on top of enforcing $q_{A \cup B} = q_A q_B$ for X_A disconnected from X_B , we have the additional constraints

$$q_A = \prod_{A_C \in \mathcal{C}(A)} q_{A_C} \quad (4)$$

for each connected set X_A , where $\mathcal{C}(A)$ are the maximal cliques in the subgraph obtained by keeping only the vertices X_A and the corresponding edges from \mathcal{G}^2 .

As a framework for the construction of bi-directed models, CDNs have three major desirable features. Firstly, the number of parameters grows with the size of the largest clique, instead of $|X_V|$. Secondly, parameters in different cliques are variation independent, since (2) is well-defined if each individual factor is a CDF. Thirdly, this is a general framework that allows not only for binary variables, but continuous, ordinal and unbounded discrete variables as well. Finally, in graphs with low tree-widths, probability densities/masses can be computed efficiently by dynamic programming [9]. To summarize, CDNs provide a restricted family of marginal independence models, but one that has computational, statistical and modeling advantages. Depending on the application, the extra constraints are not harmful in practice, as demonstrated by [10].

3 Mixed Cumulative Distribution Models

In what follows, we will extend the CDN family to general acyclic directed mixed graphs: the *mixed* cumulative distribution network (MCDN) model. In Section 3.1, we describe a higher-level factorization of the *probability* (mass or density) *function* $P(X_V)$ involving subgraphs of \mathcal{G} . In Section 3.2, we describe cumulative distribution functions that can be used to parameterize each factor defined in Section 3.1, in the special case where no directed edges exist between members of a same subgraph. Finally, in Section 3.3, we describe the general case.

Some important notation and definitions: there are two kinds of edges in an ADMG; either $X_k \rightarrow X_j$ or $X_k \leftrightarrow X_j$. In the former case (but not the latter) we call X_k a parent of X_j . We use $pa_{\mathcal{G}}(X_A)$ to represent the parents of a set of vertices X_A in graph \mathcal{G} . For a given \mathcal{G} , $(\mathcal{G})_A$ represents the subgraph obtained by removing from \mathcal{G} any vertex *not* in set A and the respective edges; $(\mathcal{G})_{\leftrightarrow}$ is the subgraph obtained by removing all directed edges. We say that a set of nodes A in \mathcal{G} is an *ancestral set* if it is closed under the ancestral relationship: if $X_v \in A$, then all ancestors of X_v in \mathcal{G} are also in A . Finally, define the districts of a graph \mathcal{G} as the connected components of $(\mathcal{G})_{\leftrightarrow}$. Hence each district is a set of vertices, X_D , such that if X_i and X_j are in X_D then there is a path connecting X_i and X_j composed entirely of bi-directed edges. Note that trivial districts are permitted, where $X_D = \{X_i\}$. Associated with each district X_{D_i} is a subgraph \mathcal{G}_i consisting of nodes $X_{D_i} \cup pa_{\mathcal{G}}(X_{D_i})$. The edges of \mathcal{G}_i are all of the edges of $(\mathcal{G})_{X_{D_i} \cup pa_{\mathcal{G}}(X_{D_i})}$ excluding all edges among $pa_{\mathcal{G}}(X_{D_i}) \setminus X_{D_i}$. Two examples are shown in Figure 2.

¹[9] describe the model in terms of factor graphs, but for our purposes a bi-directed representation is more appropriate.

²This property was called *min-independence* in [8].

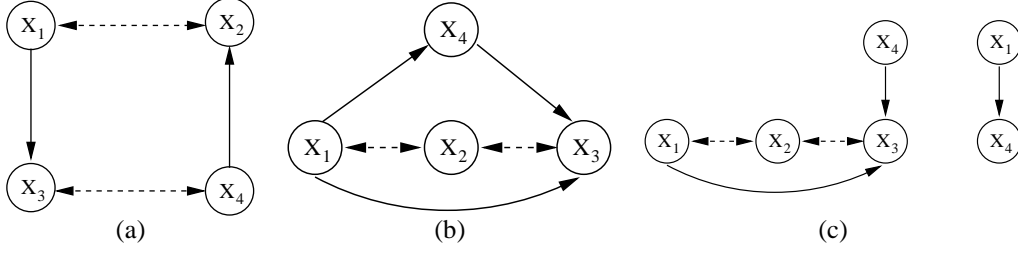


Figure 2: (a) The ADMG has two districts, $X_{D_1} = \{X_1, X_2\}$ with singleton parent X_4 , and $X_{D_2} = \{X_3, X_4\}$ with parent X_1 . (b) A more complicated example with two districts. Notice that the district given by $X_{D_1} = \{X_1, X_2, X_3\}$ has as external parent X_4 , but internally some members of the district might be parents of other members. The other district is a singleton, $X_{D_2} = \{X_4\}$. (c) The two corresponding subgraphs \mathcal{G}_1 and \mathcal{G}_2 are shown here.

3.1 District factorization

Given any ADMG \mathcal{G} with vertex set X_V , we parameterize its probability mass/density function as:

$$P(X_V) = \prod_{i=1}^K P_i(X_{D_i} \mid pa_{\mathcal{G}}(X_{D_i}) \setminus X_{D_i}) \quad (5)$$

where $\{X_{D_1}, X_{D_2}, \dots, X_{D_K}\}$ is the set of districts of \mathcal{G} . That is, each factor is a probability (mass/density) function for X_{D_i} given its set of parents in \mathcal{G} (that are not already in X_{D_i}). We require that

- Each $P_i(X_{D_i} \mid pa_{\mathcal{G}}(X_{D_i}) \setminus X_{D_i})$ is Markov with respect to \mathcal{G}_i ,

where a probability function $P(\cdot)$ is *Markov with respect* to a ADMG \mathcal{G} if any conditional independence constraint encoded in \mathcal{G} is exhibited in $P(\cdot)$.

The relevance of this factorization is summarized by the following result.

Proposition 1. *A probability function $P(X_V)$ is Markov with respect to \mathcal{G} if it can be factorized according to (5) and each $P_i(X_{D_i} \mid pa_{\mathcal{G}}(X_{D_i}) \setminus X_{D_i})$ is Markov with respect to the respective \mathcal{G}_i .*

Proofs of all results are in Appendix A.

Note that (5) is seemingly cyclical: for instance, Figure 2(a) implies the factorization $P_1(X_1, X_2 \mid X_4)P_2(X_3, X_4 \mid X_1)$. This suggests that there are additional constraints tying parameters across different factors. However, there are no such constraints, as guaranteed through the following result:

Proposition 2. *Given an ADMG \mathcal{G} with respective subgraphs $\{\mathcal{G}_i\}$ and districts $\{X_{D_i}\}$, any collection of probability functions $P_i(X_{D_i} \mid pa_{\mathcal{G}}(X_{D_i}) \setminus X_{D_i})$, Markov with respect to the respective \mathcal{G}_i , implies that (5) is a valid probability function (a non-negative function that integrates to 1).*

The implication is that one can independently parameterize each individual $P_i(\cdot \mid \cdot)$ to obtain a valid $P(X_V)$ Markov with respect to any given ADMG \mathcal{G} . In the next sections, we show how to parameterize each $P_i(\cdot \mid \cdot)$ by factorizing its corresponding cumulative distribution function.

3.2 Models with barren districts

Consider first the case where district X_{D_i} is *barren*, that is, no $X_v \in X_{D_i}$ has a parent also in X_{D_i} [20]. For a given \mathcal{G}_i with respective district X_{D_i} , consider the following function:

$$F_i(x_{D_i} \mid pa_{\mathcal{G}}(X_{D_i})) \equiv \left[\prod_{X_S \in \mathcal{C}_i} F_S(x_S \mid pa_{\mathcal{G}}(X_{D_i})) \right] \left[\prod_{X_v \in X_{D_i}} F_v(x_v \mid pa_{\mathcal{G}}(X_v)) \right] \quad (6)$$

where \mathcal{C}_i is the set of cliques in $(\mathcal{G}_i)_{\leftrightarrow}$. Each term on the right hand side is a conditional cumulative distribution function: for sets of random variables Y and Z , $F(y \mid z) \equiv P(Y \leq y \mid Z = z)$.

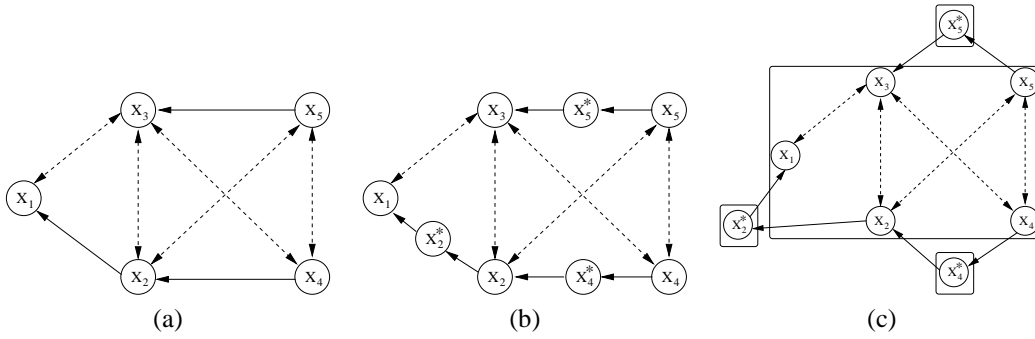


Figure 3: (a) A mixed graph with a single district that includes all five vertices. (b) The modified graph after including artificial vertices (artificial vertices for childless variables are ignored). (c) A display of the four districts of the modified graph in individual boxes. All districts are now barren, i.e., no directed edges can be found within a district.

Proposition 3. $F_i(x_{D_i})$ is a CDF for any choice of $\{\{F_S(x_S)\}, \{F_v(x_v | \text{pa}_{\mathcal{G}}(X_v))\}\}$. If, according to each $F_S(x_S)$, $X_s \in X_S$ is marginally independent of any element in $\text{pa}_{\mathcal{G}}(X_{D_i}) \setminus \text{pa}_{\mathcal{G}}(X_s)$, the corresponding conditional probability function $F_i(x_{D_i} | \text{pa}_{\mathcal{G}}(X_{D_i}))$ is Markov with respect to \mathcal{G}_i .

Notice that the structure of type IV chain graphs [3] is a special case of ADMGs with barren districts. The parameterization of [3] is complete for such graph models, but requires exponentially many parameters even in sparse models.

To obtain the probability function (5), we calculate each $P_i(X_{D_i} | \text{pa}_{\mathcal{G}}(X_{D_i}) \setminus X_{D_i})$ by differentiating the corresponding (6) with respect to X_{D_i} . Although this operation, in the discrete case, is in the worst-case exponential in $|X_{D_i}|$, it can be performed efficiently for graphs where $(\mathcal{G})_{\leftrightarrow}$ has low tree-width [9].

3.3 The general case: reduction to barren case

We reduce graphs with general districts to graphs with only barren districts by introducing artificial vertices. Create a graph \mathcal{G}^* with the same vertex set as \mathcal{G} and the same bi-directed edges. For each vertex X_v in \mathcal{G} , perform the following operation:

- add an artificial vertex X_v^* to \mathcal{G}^* ;
- add the edge $X_v \rightarrow X_v^*$ to \mathcal{G}^* , and make the children of X_v^* to be the original children of X_v in \mathcal{G} ;
- define the model $P(X_V, X_V^*)$ to have the same factors (5) as $P(X_V)$, but substituting every occurrence of X_v in $\text{pa}_{\mathcal{G}}(X_{D_i})$ by the corresponding $\text{pa}_{\mathcal{G}^*}(X_{D_i})$. Moreover, define $P_v^*(X_v^* | X_v)$ such that

$$P_v^*(X_v^* = x | X_v = x) = 1 \quad (7)$$

$$P(X_V, X_V^*) = \prod_{i=1}^K P_i(X_{D_i} | \text{pa}_{\mathcal{G}^*}(X_{D_i}) \setminus X_{D_i}) \prod_{X_v \in X_V} P_v^*(X_v^* | X_v) \quad (8)$$

Since the last group of factors is identically equal to 1, they can be dropped from the expression.

From (7), it follows that $P(X_V = x_V, X_V^* = x_V) = P(X_V = x_V)$. Since no two vertices in the same district can now have a parent-child relation, all districts in \mathcal{G}^* are barren and as such we can parameterize $P(X_V = x_V, X_V^* = x_V)$ according to the results of the previous section. A similar trick was exploited by [24] to reduce a problem of modeling ADMG probit models to Gaussian models.

Figure 3 provides an example, adapted from [20]. The graph has a single district containing all vertices. The corresponding transformed graph generates several singleton districts composed of one artificial variable either. In Figure 3(c), we rearrange such districts to illustrate the decomposition described in Section 3.1.

4 Copula MCDNs and Parameter Estimation

The main result of Section 3 is that we can parameterize a MCDN model by parameterizing the factors in Equation (6) corresponding to each district, which are then tied together by the joint model (8). However, we have not yet specified how to construct each F_S and F_v . In this section, we describe a particularly convenient way of parameterizing such factors. We introduce *copula MCDN models* – a particular instantiation of the MCDN family – and how to estimate its parameters.

Copulas are a flexible approach to defining dependence among a set random variables. This is done by specifying the dependence structure and the marginal distributions separately [16] (see also [13] for a machine learning perspective). Simply put, a copula function $C(u_1, \dots, u_t)$ is just the CDF of a set of dependent random variables, each with the uniform marginal distribution over $[0, 1]$. To define a joint distribution over a set of variables $\{X_v\}$ with arbitrary marginal CDFs $F_v(x_v)$, we simply transform each X_v into a uniform variable u_v over $[0, 1]$ using $u_v \equiv F_v(x_v)$. The resulting joint CDF $F(x_1, \dots, x_t) = C(F_1(x_1), \dots, F_t(x_t))$ incorporates both the dependence encoded in C and the marginal distributions F_v .

Returning to ADMGs, let \mathcal{G}_i be the subgraph corresponding to a barren district X_{D_i} . We parameterize a conditional CDF $F_i(x_{D_i} | pa_{\mathcal{G}}(X_{D_i}))$ of form (6) Markov with respect to \mathcal{G}_i by defining the marginal CDFs and copula dependence separately. In our implementation the marginal probability for binary or ordinal X_v is an unconstrained conditional probability mass function. The ordering over the values of X_v , \preceq , naturally defines the marginal $F_v(x_v | pa_{\mathcal{G}}(X_v))$:

$$F_v(x_v | pa_{\mathcal{G}}(X_v)) = \sum_{x \preceq x_v} \eta_x^{pa_{\mathcal{G}}(X_v)} \quad (9)$$

where η are the marginal parameters; conditioned upon the parents of X_v , $\eta_x^{pa_{\mathcal{G}}(X_v)}$ is simply the probability that $X_v = x$. In our implementation for continuous X_v , we define the marginal $F_v(x_v | pa_{\mathcal{G}}(X_v))$ using conditional Gaussians:

$$F_v(x_v | pa_{\mathcal{G}}(X_v)) = \Phi(x_v; \sum_{j=1}^K \eta_{vj} \phi_j(pa_{\mathcal{G}}(X_v)), \sigma_v^2), \quad (10)$$

with variance σ_v^2 and mean given by a linear regressor of fixed basis functions $\phi_j(\cdot)$.

For a copula with the required bi-directed dependence among X_{D_i} , we adopt the approach of product copulas [15]. For each clique S in \mathcal{G}_i let $C_S(u_S)$ be a $|S|$ -dimensional copula. Let d_v be the number of cliques variable X_v is in and define $a_v \equiv u_v^{1/(d_v+1)}$. The product of copulas given by:

$$C_{D_i}(u_{D_i}) = \prod_{S \in \mathcal{C}_i} C_S(a_S) \prod_{v \in D_i} a_v \quad (11)$$

can be shown to be a copula itself [15]. Plugging in the marginal distributions by defining $u_v \equiv F_v(x_v | pa_{\mathcal{G}}(X_{D_i}))$, the joint CDF over x_{D_i} becomes:

$$F_i(x_{D_i} | pa_{\mathcal{G}}(X_{D_i})) = \left[\prod_{S \in \mathcal{C}_i} C_S(a_S) \right] \left[\prod_{v \in D_i} a_v \right] \quad \text{where } a_v \equiv F_v(x_v | pa_{\mathcal{G}}(X_v))^{1/(d_v+1)} \quad (12)$$

The joint CDF has the form (6) required to be Markov with respect to \mathcal{G}_i .

We take an easy approach to parameter estimation commonly employed in the copula literature:

1. fit the (conditional) marginals in (9) or (10) individually (by maximizing likelihood);
2. calculate the corresponding “pseudodata” a_v ;
3. plug the estimated “pseudodata” into (12), and maximize the likelihood of the product copula (12). Note that information from the parents has been absorbed into the calculation of a_v via (9) or (10).

Although the result is not a maximum likelihood estimator, it is a practical procedure that does give consistent estimators [13]. Given the pseudodata, the third step is maximum likelihood estimation of a CDN model as discussed by [10]. In our implementation, used in Section 5, we substitute Step 3 by something even simpler to program³, while providing a proof of concept for the feasibility of Bayesian procedures: we put a prior over the copula parameters and do Metropolis-Hastings (MH) with a Gaussian random walk proposal. To calculate the MH ratio we only need the likelihood function, which again can be obtained from the message-passing scheme of [9, 10].

³Maximum likelihood estimation requires the gradient of the density with respect to the parameters. That is, we need derivatives on top of the message-passing scheme that transforms a CDF into a density function [10].

Table 1: Average difference in log predictive per observations (in millibits) and standard errors. $\# \mathbf{v}$ is the number of variables and $\# \mathbf{b}$ is the number of bi-directed edges in the ADMG.

Data set	$\# \mathbf{v}$	$\# \mathbf{b}$	Variables marginalized out	$\mathbb{E} [\Delta_{\text{DAG}}] \pm \text{s.e.}$
Insurance	25	2	Driving Skill, Mileage	72.72 ± 17.15
Alarm	33	4	Err Cauter, TPR, KinkedTube, ArtCO2	76.27 ± 13.78
BreastCancer	10	5	<i>structure inferred</i>	686.50 ± 76.62
SPECTF	44	25	<i>structure inferred</i>	-21.14 ± 25.74

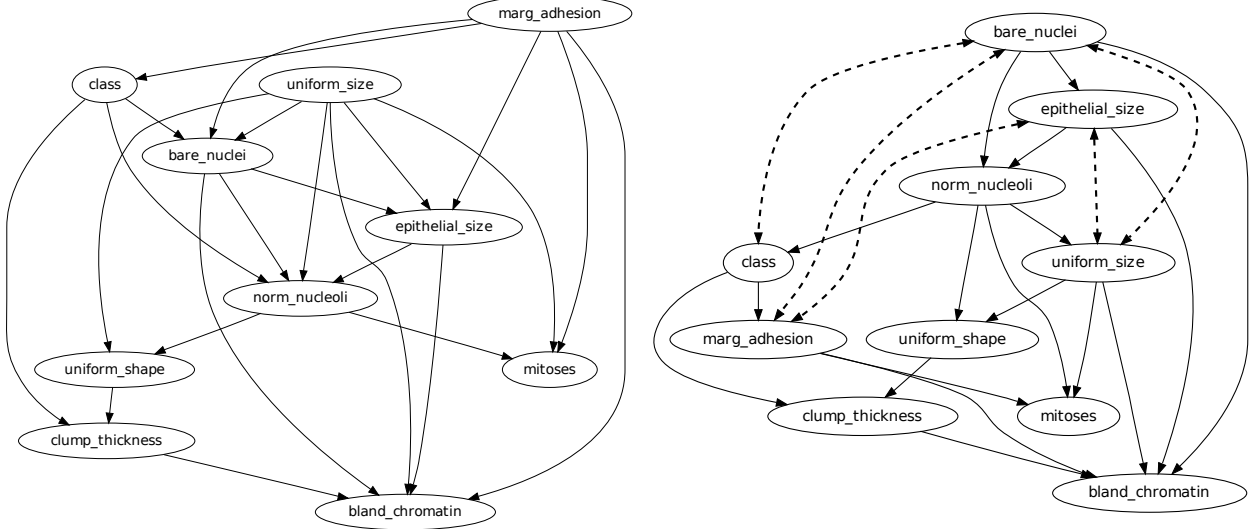


Figure 4: DAG (left) and ADMG (right) structures inferred from the Wisconsin breast cancer data set.

5 Experiments

We evaluate the usefulness of the MCDN formalism by comparing the K -fold cross validated log-predictive probabilities of copula MCDNs and DAGs on four data sets. Two data sets are synthetic (from the alarm [1] and insurance networks [11]) so that the ground truth structure is known and we can compare against an overparameterized DAG. The non-synthetic data sets are both from the UCI repository (the Wisconsin breast cancer and SPECTF data sets [26, 6]). All data sets, except for the SPECTF data set which is continuous, consist of ordinal or binary variables.

In our experiments, copula MCDNs are parameterized as described in (9) or (10), and (12). We use Frank copulas, for computational convenience, with Gaussian $\mathcal{N}(0, 10)$ priors on their parameters θ .

Known structure Several common cause variables (listed in table 1) were marginalized out of the data to introduce bi-directed edges to the true structure. An overparameterized DAG is able, parametrically, to capture a broader set of conditional dependencies (by having additional edges as well as broader parameterization) than those of a copula MCDN; however it has many more parameters (exponential in the parents of the district of the corresponding MCDN). Hence we compare these models on a small sample size of 300.

The difference, in millibits, of the log predictive probability between that of the copula MCDN and of the overparameterized DAG, per cross-validation test set, is calculated as follows:

$$\Delta_{\text{DAG}} = \frac{1000}{n_k} [\log_2 \tilde{p}(x_k | \mathcal{D}_k, \eta_k, \text{MCDN}) - \log_2 p(x_k | \mathcal{D}_k, \eta_k, \text{DAG})]$$

where x_k and \mathcal{D}_k are the k th test and training set, respectively, and η_k are the maximum likelihood parameters of the marginals from \mathcal{D}_k .

We calculate the predictive probability of the data set, $\tilde{p}(x_k | \mathcal{D}_k, \eta_k, \text{MCDN})$, by averaging $p(x_k | \mathcal{D}_k, \eta_k, \theta, \text{MCDN})$ over samples of the copula parameter θ . Positive Δ_{DAG} tells us on average how many millibits better the prediction from the MCDN is over the DAG model. In both cases the log predictive probabilities were significantly higher, although slight. Comparing to a DAG with marginal parameters marginalized produced the same numbers (up to 5 s.f.) shown in table 1.

Unknown structure Next we ran an experiment on ordinal data without known structure. We used the original Wisconsin breast cancer data set from the UCI repository [26]. The ADMG and DAG structures shown in figure 4 were inferred using MBCS* [18] and the χ^2 test. We then repeated the procedure described above, instead calculating Δ_{DAG} relative to the inferred DAG rather than an overparameterized DAG, to obtain the results also shown in table 1. On average, the model performed encouragingly.

Finally, we used the SPECTF continuous data set from the UCI repository [6]. We used this data in a more realistic fashion: instead of learning the structure from the entire data set then performing predictions of subsets, the structure learning is incorporated into the K -fold cross validation. We used $K = 5$ for this experiment and a score-based structure learning algorithm [22] to find the DAG followed by fitting the bi-directed edges using the residuals with the directed structure fixed. Furthermore, if districts were not tree-structures, they were thinned into trees (ordered by weakest residuals). The residuals were fit by testing marginal independence using [7]. This combined technique allowed the structure to be inferred efficiently.

We compared this copula MCDN to a Gaussian DAG model (fit using just the DAG learning algorithm of [22] and maximum likelihood). The results are shown in table 1. The number of bi-directed edges given is the average over the $K = 5$ cross validation folds.

In this case, the copula MCDN performed worse than the DAG model. Note that the fitting procedure is suboptimal for MCDNs and, for computational efficiency, does not alternate between learning directed and bi-directed edges, and the bi-directed structure is limited to tree-structured. We also tried fitting a copula CDN, that is, omitting the DAG search step and just fitting the residuals. Compared to this model, the MCDN had an average difference of $11,504 \pm 2,456$ millibits suggesting that the DAG marginals are dominating the copula MCDN fit on these data.

6 Conclusion

Acyclic directed mixed graphs are a natural generalization of DAGs. While ADMGs date back at least to [27], the potential of this framework has only recently been translated into practical applications due to advances into complete parameterizations of Gaussian and discrete networks [21, 5, 20]. The framework of cumulative distribution networks [9, 10] introduced new approaches for more constrained by widely applicable families of marginal independence (bi-directed) models. By extending CDNs to the full ADMG case, we expect that ADMGs will be readily accessible and as widespread as DAG models.

There are several directions for future work. While classical approaches for learning Markov equivalence classes of ADMGs have been developed by means of multiple hypothesis tests of conditional independencies [25], a model-based approach based on Bayesian or penalized likelihood functions can deliver more robust learning procedures and a more natural way of combining data with structural prior knowledge. ADMG structures can also play a role in multivariate supervised learning, that is, structured prediction problems. For instance, [23] introduced some simple models for relational classification inspired by ADMG models and by the link to seemingly unrelated regression [28]. However, efficient ADMG-structured prediction methods and new advanced structural learning procedures will need to be developed.

References

- [1] J. Binder, D. Koller, S. Russell, and K. Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29:213–244, 1997.
- [2] K. Bollen. *Structural Equations with Latent Variables*. John Wiley & Sons, 1989.
- [3] M. Drton. Discrete chain graph models. *Bernoulli*, 15:736–753, 2009.
- [4] M. Drton and T. Richardson. A new algorithm for maximum likelihood estimation in Gaussian models for marginal independence. *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, 2003.
- [5] M. Drton and T. Richardson. Binary models for marginal independence. *Journal of the Royal Statistical Society, Series B*, 70:287–309, 2008.
- [6] A. Frank and A. Asuncion. UCI machine learning repository, 2010.
- [7] A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. *Neural Information Processing Systems*, 2007.
- [8] J. Huang. *Cumulative Distribution Networks: Inference, Estimation and Applications of Graphical Models for Cumulative Distribution Functions*. PhD Thesis, University of Toronto, Department of Computer Science, 2009.
- [9] J. Huang and B. Frey. Cumulative distribution networks and the derivative-sum-product algorithm. *UAI*, 2008.

- [10] J. Huang and N. Jojic. Maximum-likelihood learning of cumulative distribution functions on graphs. *13th International Conference on Artificial Intelligence and Statistics, AISTATS*, 2010.
- [11] B. I. S. HJ, C. RM, and C. GF. The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proceedings of the 2nd European Conference on Artificial Intelligence in Medicine*, pages 247–256, 1989.
- [12] H. Joe. *Multivariate Models and Dependence Concepts*. Chapman-Hall, 1997.
- [13] S. Kirshner. Learning with tree-averaged densities and distributions. *Neural Information Processing Systems*, 2007.
- [14] S. Lauritzen. *Graphical Models*. Oxford University Press, 1996.
- [15] E. Liebscher. Construction of asymmetric multivariate copulas. *Journal of Multivariate Analysis*, 99:2234–2250, 2008.
- [16] R. Nelsen. *An Introduction to Copulas*. Springer-Verlag, 2007.
- [17] J. Pearl. *Probabilistic Reasoning in Expert Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.
- [18] J.-P. Pellet. Finding latent causes in causal networks: an efficient approach based on Markov blankets. *Neural Information Processing Systems*, 2008.
- [19] T. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30:145–157, 2003.
- [20] T. Richardson. A factorization criterion for acyclic directed mixed graphs. *Proceedings of the 25th International Conference on Uncertainty in Artificial Intelligence*, 2009.
- [21] T. Richardson and P. Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30:962–1030, 2002.
- [22] M. Schmidt, A. Niculescu-Mizil, and K. Murphy. Learning graphical model structure using l1-regularization paths. *AAAI’07*, 2007.
- [23] R. Silva, W. Chu, and Z. Ghahramani. Hidden common cause relations in relational learning. *Neural Information Processing Systems (NIPS ’07)*, 2007.
- [24] R. Silva and Z. Ghahramani. The hidden life of latent variables: Bayesian learning with mixed graph models. *Journal of Machine Learning Research*, to appear, 2009.
- [25] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. Cambridge University Press, 2000.
- [26] W. Wolberg and O. Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In *Proceedings of the National Academy of Sciences*, volume 87, pages 9193–9196, 1990.
- [27] S. Wright. Correlation and causation. *Journal of Agricultural Research*, pages 557–585, 1921.
- [28] A. Zellner. An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *Journal of the American Statistical Association*, 1962.
- [29] J. Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9:1437–1474, 2008.

APPENDIX A – PROOFS

Proposition 1. A probability function $P(X_V)$ is Markov with respect to \mathcal{G} if it can be factorized according to (5), given that each $P_F(X_{D_i} \mid pa_{\mathcal{G}}(X_{D_i}) \setminus X_{D_i})$ is Markov with respect to \mathcal{G}_i .

Before we prove this theorem, we need to state the following result from [19]. Given an ancestral set A , the *Markov blanket* of vertex X_v in A , $mb(X_v, A)$, is given by the district of X_v in $(\mathcal{G})_A$ (except X_v itself) along with all parents of elements of this district. Let a *total ordering* \prec of the vertices of \mathcal{G} be any ordering such that if $X_v \prec X_t$, then X_t is not an ancestor of X_v in \mathcal{G} . A probability measure is said to satisfy the *ordered local Markov condition* for \mathcal{G} with respect to \prec if, for any X_v and ancestral set A such that $X_t \in A \setminus \{X_v\} \Rightarrow X_t \prec X_v$, we have X_v is independent of $A \setminus (mb(X_v, A) \cup \{X_v\})$ given $mb(X_v, A)$. The main result from [19] states:

Theorem 1. The ordered local Markov condition is equivalent to the global Markov condition in ADMGs⁴.

Proof of Proposition 1: The proof is done by induction on $|X_V|$, with the case $|X_V| = 1$ being trivial. We will show that if $P(X_V)$ is a probability function that factorizes according to (5), as given by an ADMG \mathcal{G} , then $P(X_V)$ is Markov with respect to \mathcal{G} . To prove this, first notice there must be some X_v with no children in \mathcal{G} , since the graph is acyclic. Let X_{D_i} be the district of X_v . By assumption,

$$\begin{aligned} P(X_V) &= P_F(X_v \mid X_{D_i} \cup pa_{\mathcal{G}}(X_{D_i})) \times P_F(X_{D_i} \setminus X_v \mid pa_{\mathcal{G}}(X_{D_i}) \setminus X_{D_i}) \\ &\times \prod_{j \neq i} P_F(X_{D_j} \mid pa_{\mathcal{G}}(X_{D_j}) \setminus X_{D_j}) \end{aligned} \quad (13)$$

Since X_v is childless, it does not appear in any of the factors in the expression above, except for the first. Hence,

$$P(X_V \setminus X_v) = P_F(X_{D_i} \setminus X_v \mid pa_{\mathcal{G}}(X_{D_i}) \setminus X_{D_i}) \times \prod_{j \neq i} P_F(X_{D_j} \mid pa_{\mathcal{G}}(X_{D_j}) \setminus X_{D_j}) \quad (14)$$

⁴Notice this reduces to the standard notion of local independence in DAGs, where a vertex is independent of its (non-parental) non-descendants given its parents, from which d-separation statements can be derived [14, 17].

which by induction hypothesis is Markov with respect to the marginal graph $(\mathcal{G})_{X_V \setminus X_v}$ (one minor detail is that $(\mathcal{G})_{X_V \setminus X_v}$ might have more districts than \mathcal{G} after removing X_v . However, the result still holds by further factorizing $P_F(X_{D_i} \setminus X_v \mid pa_{\mathcal{G}}(X_{D_i}) \setminus X_{D_i})$ according to the newly formed districts of $X_{D_i} \setminus X_v$ – which is possible by the construction of $P_F(\cdot)$ and \mathcal{G}_i). By the ordered local Markov property for ADMGs and any ordering \prec where X_v is the last vertex, probability function $P(X_V)$ will be Markov with respect to \mathcal{G} if, according to $P(X_v)$, the Markov blanket of X_v in \mathcal{G} makes X_v independent of the remaining vertices. But this true by construction, since this Markov blanket is contained in $X_{D_i} \cup pa_{\mathcal{G}}(X_{D_i})$ according to Theorem 1. \square

Notice that factorization (5) is seemingly cyclical: for instance, Figure 2(a) implies the factorization $P_F(X_1, X_2 \mid X_4)P_F(X_3, X_4 \mid X_1)$. This suggests that there are additional constraints tying parameters across different factors. However, there are no such constraints, as guaranteed through the following result:

Proposition 2. *Given an ADMG \mathcal{G} with respective subgraphs $\{\mathcal{G}_i\}$ and districts $\{X_{D_i}\}$, any collection of probability functions $P_F(X_{D_i} \mid pa_{\mathcal{G}}(X_{D_i}) \setminus X_{D_i})$, Markov with respect to the respective \mathcal{G}_i , implies that (5) is a valid probability function (a non-negative function that integrates to 1).*

Proof: It is clear that (5) is non-negative. We have to show it integrates to 1. As in the proof of Proposition 1, first notice there must be some X_v with no children in \mathcal{G} , since the graph is acyclic. Those childless vertices can be marginalized as in Equation (14) if they do not appear on the right-hand side of any factor $P_F(\cdot \mid \cdot)$, and removed from the graph along with all edges adjacent to them. After some marginalizations, suppose that in the current marginalized graph, a childless vertex X_\emptyset appears on the right-hand side of some factor $P_F(X_{D_i} \mid pa_{\mathcal{G}}(X_{D_i}) \setminus X_{D_i})$. Because X_\emptyset has no children in X_{D_i} , by construction X_{D_i} and X_\emptyset are independent given the remaining elements in $pa_{\mathcal{G}}(X_{D_i}) \setminus X_{D_i}$. As such, X_\emptyset can be removed from the right-hand side of all remaining factors, and then marginalized. The process is repeated until the last remaining vertex is marginalized, giving 1 as the result. \square

Proposition 3. *$F_i(x_{D_i} \mid pa_{\mathcal{G}}(X_{D_i}))$ is a CDF for any choice of $\{\{F_S(x_S \mid pa_{\mathcal{G}}(X_S))\}, \{F_v(x_v \mid pa_{\mathcal{G}}(X_v))\}\}$. If, according to each $F_S(x_S \mid \cdot)$, $X_S \in X_S$ is independent of any element in $pa_{\mathcal{G}}(X_{D_i}) \setminus pa_{\mathcal{G}}(X_S)$, the corresponding conditional probability function $F_i(x_{D_i} \mid pa_{\mathcal{G}}(X_{D_i}))$ is Markov with respect to \mathcal{G}_i .*

Proof: Each factor in (6) is a CDF with respect to X_{D_i} , with $pa_{\mathcal{G}}(X_{D_i})$ fixed, and hence its product is also a CDF [9]. To show the Markov property, it is enough to consider the modified graph \mathcal{G}'_i constructed by transforming all directed edges in \mathcal{G}_i into bi-directed edges, since the implied distributions conditional on $pa_{\mathcal{G}}(X_{D_i})$ for \mathcal{G}'_i and \mathcal{G}_i are Markov equivalent [21]. It follows directly from the assumptions and the properties of CDFs that disconnected sets in \mathcal{G}'_i are marginally independent, which corresponds to the Markov properties of bi-directed graph \mathcal{G}'_i [19]. \square

APPENDIX B – BINARY CASE: RELATION TO COMPLETE PARAMETERIZATION

A complete parameterization for binary ADMG models is described by [20]. As we will see, parameters are defined in the context of different marginals, analogous to the purely bi-directed case [5].

As in the bi-directed case, the joint probability distribution is given by an inclusion-exclusion scheme:

$$P(X_V = \alpha(V)) = \sum_{C: \alpha^{-1}(0) \subseteq C \subseteq V} (-1)^{|C \setminus \alpha^{-1}(0)|} \prod_{H \in [C]_{\mathcal{G}}} P(X_H = 0 \mid X_{tail(H)} = \alpha(tail(H))) \quad (15)$$

where $\alpha(V)$ is a binary vector in $\{0, 1\}^{|X_V|}$ and $\alpha^{-1}(0)$ is a function that indicates which elements in X_V were assigned to be zero.

Each C indicates which elements are set to zero in the respective term of the summation. Depending on C , the factorization changes. $[C]_{\mathcal{G}}$ is a set of subsets of X_V : one subset per district, each subset being barren in \mathcal{G} . The corresponding $tail(H)$ is the Markov blanket for the ancestral set that contains H as its set of childless vertices.

As in our discussion of standard CDNs, Equation (15) can be interpreted as the CDF-to-probability transformation (3). It can be rewritten as

$$P(X_V = \alpha(V)) = \sum_{C: \alpha^{-1}(0) \subseteq C \subseteq V} (-1)^{|C \setminus \alpha^{-1}(0)|} \times \prod_{H \in D_i \cap [C]_{\mathcal{G}}} P(X_{D_i} \setminus tail(H) \leq \alpha(V) \mid X_{tail(H)} = \alpha(tail(H)))$$

Hence, this parameterization can also be interpreted as a CDF parameterization. One important difference is that each term in the summation uses only a subset of each district, $X_{D_i} \setminus tail(H)$ instead of X_{D_i} . Notice that some elements of X_{D_i} appear in the conditioning set (i.e., $tail(H)$ contains some of the remaining elements of X_{D_i} , on top of the respective parents).

The need for using subsets comes from the necessity of enforcing independence constraints entailed by bi-directed paths. As in the CDN model, the MCDN criterion factorizes each CDF according to its cliques as an indirect way of accounting for such constraints. Hence, we do not construct factorizations for different marginals: each factor within a summation term in (15) includes

all elements of each district. We enforce that they remain barren by the transformation in Section 3.3 – which is unnecessary in [20] because only barren subsets are being considered.

To understand how the parameterizations coincide, or which constraints analogous to (4) emerge in our parameterization, consider first the following example. Using the results from [20], the graph in Figure 2(a) needs the specification of the following marginals:

$$\begin{aligned}
P(X_1, X_4) &= P(X_1)P(X_4) \\
P(X_1, X_3, X_4) &= P(X_3, X_4 | X_1)P(X_1) \\
P(X_1, X_2, X_4) &= P(X_1, X_2 | X_4)P(X_4) \\
P(X_1, X_2, X_3, X_4) &= P(X_1, X_2 | X_4)P(X_3, X_4 | X_1) \\
P(X_1, X_3) &= P(X_3 | X_1)P(X_1) \\
P(X_2, X_4) &= P(X_2 | X_4)P(X_4)
\end{aligned} \tag{16}$$

As an example, the probability $P(X_{14} = 0, X_{23} = 1) \equiv P(X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 0)$ can be derived from the above factorizations and (15) as

$$\begin{aligned}
&P(X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 0) \\
&= P(X_1 \leq 0, X_2 \leq 1, X_3 \leq 1, X_4 \leq 0) - P(X_1 \leq 0, X_2 \leq 1, X_3 \leq 0, X_4 \leq 0) - \\
&\quad P(X_1 \leq 0, X_2 \leq 0, X_3 \leq 1, X_4 \leq 0) + P(X_1 \leq 0, X_2 \leq 0, X_3 \leq 0, X_4 \leq 0) \\
&= P(X_1 = 0, X_4 = 0) - P(X_1 = 0, X_3 = 0, X_4 = 0) - \\
&\quad P(X_1 = 0, X_2 = 0, X_4 = 0) + P(X_1 = 0, X_2 = 0, X_3 = 0, X_4 = 0) \\
&= P(X_1 = 0)P(X_4 = 0) - P(X_{34} = 0 | X_1 = 0)P(X_1 = 0) - \\
&\quad P(X_{12} = 0 | X_4 = 0)P(X_4 = 0) + P(X_{12} = 0 | X_4 = 0)P(X_{34} = 0 | X_1 = 0)
\end{aligned}$$

where the last line comes from the pool of possible factorizations (16). The corresponding probability using the MCDN parameterization is

$$\begin{aligned}
&= P(X_1 = 0, X_2 = 1 | X_4 = 0)P(X_3 = 1, X_4 = 0 | X_1 = 0) \\
&= (P(X_1 \leq 0, X_2 \leq 1 | X_4 = 0) - P(X_1 \leq 0, X_2 \leq 0 | X_4 = 0)) \times \\
&\quad (P(X_3 \leq 1, X_4 \leq 0 | X_1 = 0) - P(X_3 \leq 0, X_4 \leq 0 | X_1 = 0)) \\
&= (P(X_1 = 0 | X_4 = 0) - P(X_1 = 0, X_2 = 0 | X_4 = 0)) \times \\
&\quad (P(X_4 = 0 | X_1 = 0) - P(X_3 = 0, X_4 = 0 | X_1 = 0)) \\
&= (P(X_1 = 0) - P(X_1 = 0, X_2 = 0 | X_4 = 0)) \times \\
&\quad (P(X_4 = 0) - P(X_3 = 0, X_4 = 0 | X_1 = 0)) \\
&= P(X_1 = 0)P(X_4 = 0) - P(X_{34} = 0 | X_1 = 0)P(X_1 = 0) - \\
&\quad P(X_{12} = 0 | X_4 = 0)P(X_4 = 0) + P(X_{12} = 0 | X_4 = 0)P(X_{34} = 0 | X_1 = 0)
\end{aligned}$$

where the first line comes from the factorization of $P(X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 0)$ according to (5) and the fourth line comes from the Markov properties of each \mathcal{G}_i factor. Although these parameterizations have the same high-level parameters, they still do not coincide, as shown in the next example.

For a more complicated case where an extra constraint appears in our parameterization, consider Figure 3(a). In [20], it is shown that one of the parameters of the complete parameterization is $P(X_1 = 0, X_3 = 0 | X_2 = 0, X_4 = 0, X_5 = 0)$, which reflects the fact that X_1 and X_5 are dependent given all other variables. This also true in our case, except that according to Figure 3(c), our corresponding CDF is given by

$$F(x_1 | X_2)F(x_1, x_3)F(x_2, x_3)F(x_3, x_4)F(x_4, x_5)F(x_3 | X_5)F(x_2 | X_4)$$

which, evaluated at $X_{12345} = 0$, gives

$$\begin{aligned}
&P(X_1 = 0 | X_2 = 0)P(X_1 = 0, X_3 = 0)P(X_2 = 0, X_3 = 0)P(X_3 = 0, X_4 = 0) \times \\
&\quad P(X_4 = 0, X_5 = 0)P(X_3 = 0 | X_5 = 0)P(X_2 = 0 | X_4 = 0)
\end{aligned}$$

implying that $P(X_{12345} = 0)$ factorizes as $f(X_1, X_2, X_3, X_4)g(X_2, X_3, X_4, X_5)$, the generalization to (4).